TP1 : Estimation et comparaison de courbes de survie

Le compte-rendu du TP1 ne présentera pas l'étude des données de Freireich. Il s'agira de reprendre toutes les étapes de l'analyse effectuée sur les données historiques de Freireich (parties 1 et 2) et de les appliquer aux données de sevrage tabagique `"pharmacosmoking"` (partie 3).

1 Mise en œuvre sur les Données historiques de Freireich

Les données de Freiereich constituent un exemple pionnier d'utilisation de l'analyse de survie. La première publication de ces données provient de l'article Freireich et al. (1963).

Il s'agit d'étudier l'effet du traitement sur la durée de rémission de patients atteints de leucémie aiguë.

1.1 Package "survival" : premiers pas

Le jeu de données contient 42 observations :

- **time** : time in remission (in weeks)
- **status** : event status, 1 is relapse, 0 is censored
- **treatment** : treatment group : either 'placebo' or '6-MP'
- **pair** : pair id number

On charge les librairies qui vont être utilisées dans cette première partie. On charge les données et on les met dans un format requis pour l'utilisation de fonctions R pour l'analyse de survie :

► Charger les données :

```
install.packages("bpcp")
library(bpcp)
data("leuk2")
```

► Fonction "Surv" du package survival

```
install.packages("survival")
library(survival)
with(leuk2,Surv(time,status))
```

La fonction `"Surv"` permet d'identifier la variable de durée (`time`) et l'indicatrice de l'événement d'intérêt (`status`) pour la suite des traitements statistique.

Quel est l'événement d'intérêt ici? Observez le format des données obtenues avec la fonction `Surv`.

Déterminer le taux de censure, le type de censure dans chacun des groupes de traitement.

1.2 Estimation de la fonction de survie globale

On va construire une estimation de la courbe de survie, tout groupe confondu (cela n'a pas beaucoup de sens sur le plan de l'analyse de ces données et c'est simplement pour introduire les outils) :

► Estimateur de Kaplan-Meier (option `stype=1`) :

```
fit1<-survfit(Surv(time, status) ~ 1, stype=1,data = leuk2)
plot(fit1,col="blue")
```

► Estimateur de Fleming-Harrington (option `stype=2`) :

```
fit2<-survfit(Surv(time, status) ~ 1, stype=2,data = leuk2)
plot(fit2,col="red")
```

On remarque qu'un intervalle de confiance est ajouté au graphique automatiquement et on voit que l'estimateur de Fleming-Harrington est toujours supérieur à celui de Kaplan-Meier (plus optimiste!).

On peut également obtenir l'estimation de la fonction de risque cumulé grâce à l'option `ctype=1` (estimateur de Nelson-Aalen) ou `ctype=2` (estimateur de Fleming-Harrington)

1.3 Les intervalles de confiance

Les différents type d'intervalles de confiance peuvent être construits *via* l'option `conf.type=` "plain", "log" (par défaut), "log-log", "logit" ou "arcsin".

L'option `conf.int=0.95` (par défaut) permet de régler le niveau de confiance.

► Intervalle de confiance conf.type="plain"

```
fit_plain<-survfit(Surv(time, status) ~ 1, stype=1,conf.type="plain",data = leuk2)
plot(fit_plain,col="blue",xlab="Time",ylab="Survival Probability")
summary(fit_plain)
```

► Intervalle de confiance conf.type="log"

```
fit_log<-survfit(Surv(time, status) ~ 1, stype=1,conf.type="log",data = leuk2)
plot(fit_log,col="blue",xlab="Time",ylab="Survival Probability")
summary(fit_log)
```

et de même pour "log-log", "logit" ou "arcsin".

Retrouver par le calcul pour la valeur $t = 6$ les valeurs numériques obtenues dans le summary :

- l'estimation de la survie par Kaplan-Meier $\hat{S}_{KM}(6)$
- l'écart-type estimé de $\hat{\sigma}[\hat{S}_{KM}(6)]$ par la formule de Greenwood
- les bornes de l'intervalle de confiance ponctuel de niveau 95% obtenu avec l'option `conf.type="plain"`
- les bornes de l'intervalle de confiance ponctuel de niveau 95% obtenu avec l'option `conf.type="log"`
- les bornes de l'intervalle de confiance ponctuel de niveau 95% obtenu avec l'option `conf.type="log-log"`

Comparer les résultats.

Remarque : On obtient des intervalles de confiance ponctuels c'est-à-dire valides uniquement en chaque temps correspondant à la survenue d'un événement (pour chaque saut de l'estimateur).

1.4 Les bandes de confiance

Les bandes de confiance sont valables sur tout l'intervalle. Le prix à payer est que pour un même niveau de confiance, elles sont plus larges que l'intervalle de confiance. Le package `km.ci` permet de représenter les bandes de confiance sur le graphique de la courbe de survie :

```
library(km.ci)

fit_log<-survfit(Surv(time, status)~1, stype=1,conf.type="log",data = leuk2)
plot(fit_log,col="blue",lwd=2)

fitCB1<-km.ci(fit1,method=c("hall-wellner"))
lines(fitCB1, lty=2,col="red",lwd=2)
```

Explorer les différentes méthodes et comparer les bandes de confiance obtenues.

2 Comparer la fonction de survie dans plusieurs groupes

► On souhaite comparer les fonctions de survie dans les deux groupes définis par la variable **treatment** :

```
## Comparer des groupes
## variable treatment
fit_grp<-survfit(Surv(time, status) ~ treatment,data = leuk2)
summary(fit_grp)
plot(fit_grp,col=c(1,2))
plot(fit_grp,col=c(1,2),conf.int=TRUE,mark.time=TRUE)
plot(fit_grp,col=c(1,2),conf.int=TRUE,fun="F")
plot(fit_grp,col=c(1,2),conf.int=TRUE,fun="cumhaz")
```

Compléter ces graphiques en ajoutant des légendes d'axes et légendes de graphiques.

Commenter.

Remarque : L'option `conf.int=TRUE` est nécessaire pour obtenir les intervalles de confiance en présence de plusieurs groupes.

Remarque : La fonction `km.ci` ne fonctionne pas pour plusieurs groupes. Pourtant cela serait bien utile ! Il faut récupérer les valeurs des bandes de confiance pour chaque groupe et construire le graphique "à la main".

2.1 Des graphiques plus élaborés

La fonction `ggsurvplot` du package `survminer` permet de faire des graphiques plus jolis et de nombreuses options sont possibles.

```
library(survminer)
ggsurvplot(
  fit_grp,
  data=leuk2,
  conf.int=TRUE,
  pval=TRUE,
```

```
  surv.median.line="hv",
  palette=c("blueviolet","coral"),
  ggtheme=theme_light(),
  legend.labs=list("6-MP","Placebo")
)
```

Remarque : D'autres options sont possibles : `conf.int.style="step"`, `ncensor.plot=TRUE`, `log.rank.weights="n"`.

2.2 Les tests de rangs pour comparer deux ou plusieurs groupes

Comme il est délicat de comparer des courbes survie sur la simple base du graphique avec des intervalles de confiance de surcroît ponctuel, on complète l'analyse avec des tests de rangs. La fonction `survdiff` de la librairie `survival` permet d'implémenter les tests de rangs (notamment log-rank `rho=0` et Wilcoxon `rho=1`) :

```
survdiff(Surv(time, status) ~ treatment, rho=0,data=leuk2)
```

Comparer les valeurs des statistiques de tests (ou les p-values) des tests de log-rank et Wilcoxon. Le test de Wilcoxon va rejeter plus fortement l'hypothèse nulle lorsqu'il existe une différence plus marquée entre les groupes pour les événements qui se produisent précocément. Expliquez.

L'interprétation des test de rangs s'accompagnera donc toujours de l'analyse des courbes de survie. On peut noter que le graphique produit avec la fonction `ggsurvplot` fait apparaître (par défaut) la p-value du test du log-rank de comparaison des groupes (l'option `log.rank.weights="n"` permet d'avoir la p-value du test de Wilcoxon).

3 Traitement des données Pharmacosmoking

Vous étudiez les données `"pharmacoSmoking"` disponibles dans le package `"saur"` d'un essai clinique décrit par Steinberg [1] et al. (2009). Il s'agit d'étudier l'effet sur la durée de sevrage du tabac (en jours) d'une combinaison de traitements vs le patch nicotinique seul.

► Charger les librairies qui vont être utilisées, charger les données et les mettre dans un format requis pour l'utilisation de fonctions R pour l'analyse de survie.

► Se renseigner sur le jeu de données et décrire les données le plus précisément possible : identifier la variable de durée d'intérêt, quel est le type de censure, quels est le taux de censure, quelles sont les covariables ?

► Reprendre tous les éléments de l'analyse de survie pour les données `PharmacoSmoking` :

- Comparer différents groupes définis par le traitement de substitut nicotinique, l'âge (`ageGroup2` ou `ageGroup4`) le genre ou toute autre covariable qui vous semble pertinente.
- En vous appuyant sur des graphiques et des tests de rangs, proposez une analyse de ce jeu de données et identifiez les facteurs favorisant ou non le sevrage tabagique.

1. Steinberg, M.B. Greenhaus, S. Schmelzer, A.C. Bover, M.T., Foulds, J., Hoover, D.R., and Carson, J.L. (2009) Triple-combination pharmacotherapy for medically ill smokers : A randomized trial. Annals of Internal Medicine 150, 447-454.